Molecular Structure Determination by Convex Global Underestimation of Local Energy Minima

A.T. Phillips, J.B. Rosen, V.H. Walke

ABSTRACT. The determination of a stable molecular structure can often be formulated in terms of calculating the global (or approximate global) minimum of a potential energy function. Computing the global minimum of this function is very difficult because it typically has a very large number of local minima which may grow exponentially with molecule size. The optimization method presented involves collecting a large number of conformers, each attained by finding a local minimum of the potential energy function from a random starting point. The information from these conformers is then used to form a convex quadratic global underestimating function for the potential energy of all known conformers. This underestimator is an L_1 approximation to all known local minima, and is obtained by a linear programming formulation and solution. The minimum of this underestimator is used to predict the global minimum for the function, allowing a localized conformer search to be performed based on the predicted minimum. The new set of conformers generated by the localized search serves as the basis for another quadratic underestimation step in an iterative algorithm. This algorithm has been used to determine the structures of homopolymers of length $n \leq 30$ with no sidechains. While it is estimated that there are $O(3^n)$ local minima for a chain of length n, this method requires $O(\mathrm{n}^4)$ computing time on average. It is also shown that the global minimum potential energy values lie on a concave quadratic curve for $n \leq 30$. This important property permits estimation of the minimum energy for larger molecules, and also can be used to accelerate the global minimization algorithm.

1. Introduction

An important class of difficult global minimization problems arise as an essential feature of molecular structure calculations. The determination of a stable molecular structure can often be formulated in terms of calculating the global (or approximate global) minimum of a potential energy function (see [8]). Computing the global minimum of this function is very difficult because it typically has a very large number of local minima which may grow exponentially with molecule size.

1991 Mathematics Subject Classification. Primary 65K05; Secondary 90C26, 90C90.

The first author was supported by a grant from the Minnesota Supercomputer Institute; the second author was supported by the AFOSR grant 91-0147; the third author was supported by the Trident Scholar Program at the United States Naval Academy.

One such application is the well known protein folding problem. It is widely accepted that the folded state of a protein is completely dependent on the one-dimensional linear sequence (i.e. “primary” sequence) of amino acids from which the protein is constructed: external factors, such as enzymes, present at the time of folding have no effect on the final, or native, state of the protein. This led to the formulation of the protein folding problem: given a known primary sequence of amino acids, what would be its native, or folded, state in three-dimensional space.

Recently, several successful predictions of folded protein structures have been made and announced before the experimental structures were known (see [2], [9]). While most of these have been made with a blend of a human expert’s abilities and computer assistance, fully automated methods have shown promise for producing previously unattainable accuracy.

These machine based prediction strategies attempt to lessen the reliance on experts by developing a completely computational method. Such approaches are generally based on two assumptions. First, that there *exists* a potential energy function for the protein; and second that the folded state corresponds to the structure with the lowest potential energy (minimum of the potential energy function) and is thus in a state of thermodynamic equilibrium. This view is supported by in vitro observations that proteins can successfully refold from a variety of denatured states. Evolutionary theory also supports a folded state at a global energy minimum. Protein sequences have evolved under pressure to perform certain functions, which for most known occurrences requires a stable, unique, and compact structure. Unless specifically required for a certain function, there was no biochemical need for proteins to hide their global minimum behind a large kinetic energy barrier. While kinetic blocks may occur, they should be limited to special proteins developed for certain functions (see [1]).

2. A Simplified Model for Computational Methods

Unfortunately, finding the “true” energy function of a molecular structure, if one even exists, is virtually impossible. For example, with proteins ranging in size up to 1,053 amino acids (a collagen found in tendons), exhaustive conformational searches will never be tractable. One possible way of finding the global energy minimum is to use a simplified model of the molecular structure. By using a simplified model, the complexity of the problem formulation could be reduced to an acceptable level for optimization techniques. Care must be taken, however, to insure that the error included in such an approximation does not drive the computational solution too far away from the true native state.

One possible approximation method, which can be applied to molecules that form a linear chain, represents the molecular structure as a string of beads where the position of each bead is defined by its location relative to the previous three beads in the sequence (see [4]). In this model, the chain monomers come in two forms, H (hydrophobic) and P (polar), where the H-type monomers exhibit a strong pairwise attraction, and hence the lowest free energy is obtained by those conformations with the greatest number of HH “contacts” (see [4], [10]). One significant advantage of this formulation of the folding problem is that it allows the model to take advantage of known scientific knowledge about the chemical structure of sequences of molecules. The use of knowledge such as the Ramachandran plot (see [7]), which specifies the allowable angles between consecutive amino acids in proteins, greatly simplifies the problem. This paper will examine one such method for the global optimization of a potential energy function.

3. Potential Energy Function

Molecular structure information is generally given in terms of internal molecular coordinates which consist of bond lengths, bond angles, and the mainchain backbone torsion angles. For example, a protein consists of n amino acids in a “primary” sequence a_1, a_2,..., a_n, where a_i represents the i^{th} amino acid in the sequence. For every consecutive pair of amino acids, a_{i-1} and a_i, l_i represents the distance by which they are separated. For every three consecutive amino acids, a_{i-2}, a_{i-1}, and a_i, the bond angle, θ_i, represents the position of the third amino acid with respect to the line containing the previous two amino acids. Similarly, for every four consecutive amino acids, a_{i-3}, a_{i-2}, a_{i-1}, and a_i, the torsion angle, φ_i, represents the relative position of the fourth amino acid, a_i, with respect to the plane containing the previous three amino acids (see Figure 3.1).

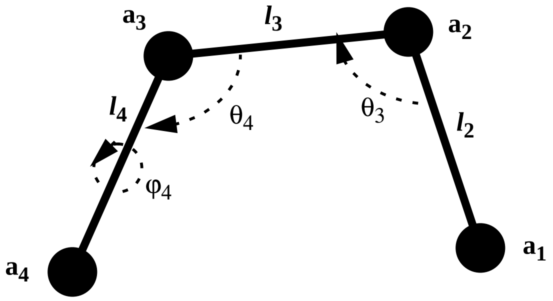

Figure 3.1 Conformation for a Four Bead Sequence

Empirical representations of the protein’s potential energy include energy terms to represent chemical bonds, angles, and torsions, as well as non-bonded interactions between amino acids farther apart in the chemical structure. This simplified definition resembles a physical model in which beads (e.g. amino acids) are connected by springs (first term of (1)) at a distance of l_i. Bond angles, θ_i, which are determined by a sequence of three beads (a_{i-2},a_{i-1},a_i) are maintained by similar “springs” (the second term of (1)). Torsion angles, φ_i, are modeled by a trigonometric based penalty (third term of (1)). Such a formulation is often called a molecular mechanics potential. One often used formula for the overall potential U is:

$$(1)\qquad U = \sum_i \frac{1}{2} k_l (l_i - l_0)^2 + \sum_i \frac{1}{2} k_\theta (\theta_i - \theta_0)^2 + \sum_i \frac{V_{\bar{n}}}{2} [1 + \cos(\bar{n}\varphi_i + \delta)] + \sum_{i,l} \varepsilon_{i,l} \left(\left(\frac{\sigma_{i,l}}{r_{i,l}} \right)^{12} - 2H_{i,l} \left(\frac{\sigma_{i,l}}{r_{i,l}} \right)^{6} \right)$$

where k_l and k_θ are the bond stretching and angle bending force constants. In the third term of the equation, V_n is the n-fold torsional constant with a phase shift of δ. This term provides for preferred torsion angles within the molecular structure. For example, if $n = 3$ and the phase shift δ is zero, preferred torsion angles (configurations resulting in a small penalty) can be found at 60˚, 180˚, and 300˚. The constants l_0 and θ_0 represent the preferred bond length and bond angle for each sequence of consecutive beads.

The fourth term of the (1), $\sum_{i,l} \left((\sigma_{i,l}/r_{i,l})^{12} - 2H_{i,l} (\sigma_{i,l}/r_{i,l})^{6} \right)$, is known as the Lennard-Jones pairwise potential (see Figure 3.2). This term defines the potential energy contributions of all beads separated by more than two along the primary chain. In (1), $\varepsilon_{i,l}$ and $\sigma_{i,l}$ are the Lennard-Jones coefficients, which are constants defined by the relationships

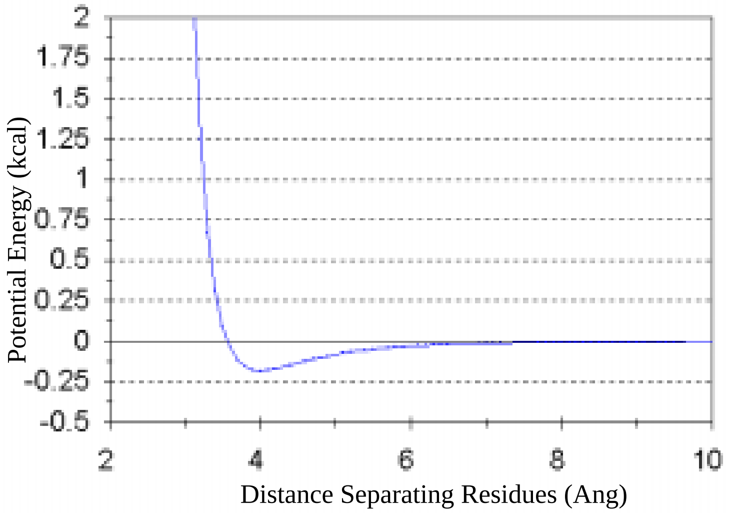

Figure 3.2 Lennard-Jones Pairwise Potential (ε=0.181, σ=4.0)

between the two specific beads (e.g. amino acids) involved. The terms involving $r_{i,l}$ in the Lennard-Jones expression represent the Euclidean distance between the beads a_i and a_l (see [3], [5], [8]). The constant $H_{i,l} = 1$ if beads a_i and a_l are both H-type (attractive monomers), and hence both a repulsive force (ensuring that the chain is "self-avoiding") and an attractive force (since the beads are H-H) are added to the potential energy. On the other hand, $H_{i,l} = 0$ if the beads a_i and a_l are H-P, P-H, or P-P pairs, so that the Lennard-Jones contribution to the total potential energy is just the repulsive force that ensures self-avoidance.

4. Coordinate Representations

Potential energy functions for molecular conformations usually involve computing Euclidean distances between molecules. For example the potential function in (1) contains the Lennard-Jones pairwise potential, $\sum_{i,l}\left((\sigma_{i,l}/r_{i,l})^{12} - 2H_{i,l}(\sigma_{i,l}/r_{i,l})^{6}\right)$, which depends on the lengths between all beads separated by more than two residues along the primary chain. Unfortunately, computing distances using bond lengths, bond angles, and torsion angles is extremely difficult. Since optimization methods require this computation to be executed often, it is desirable to convert the representation of the problem into cartesian coordinates. Once the conversion to three dimensional cartesian coordinates has been completed, the length between the residues can be found in $O(n^2)$ time. However, the conversion itself becomes very complex since every set of three residues creates *its own reference frame* for bond lengths, bond angles, and torsion angles.

To perform the conversion from internal molecular coordinates (bond lengths, bond angles, and torsion angles) to a cartesian representation, the first three beads in the sequence can be fixed, without loss of generality, as depicted in Figure 4.1. The first bead, a_1, is fixed at the origin, (0,0,0). The second bead, a_2, is positioned at $(-l_2,0,0)$, a distance from a_1 equal to the bond length, l_2, along the negative x axis. The location of the third bead, a_3, is fixed at $(l_3\cos(\theta_3)-l_2, l_3\sin(\theta_3), 0)$. The fourth and any other beads constituting the primary sequence are found using the cartesian representation associated with the pre-

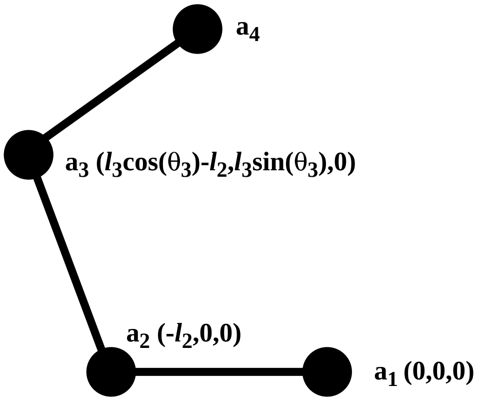

Figure 4.1 Four Bead Sequence (Cartesian Coordinates)

vious three beads, and the bond length, bond angle, and torsion angle associated with the bead under consideration. The cartesian representation of this and all following beads in the sequence can be computed in a variety of ways, two of which are presented in [11] and [12]. Appendix A (Section 12) presents a discussion of both the computations required to perform the transformations, and also a derivation of an analytic form of the gradient of the potential function U.

5. *n*-Chain Homopolymer Problem

One simple application of the "folding" problem, which can be modeled as a string of beads, is the *n-chain homopolymer problem*. Homopolymers can be modeled in manner similar to proteins: a string of beads (a_i, where $i = 1,...,n$), positioned by their respective bond lengths, l_i, bond angles, θ_i, and torsion angles, φ_i. However, each "residue" in this case can be considered *identical* (i.e. all of type H). Furthermore, it is well known that in the case of hydrocarbons, these homopolymers have "preferred" bond lengths $l_0 = 1.526$ Ang and bond angles $\theta_0 = 109.47^\circ$ (1.91 rad), and hence this information can be used in some optimization techniques to reduce the number of problem parameters.

Heptane is an n-chain hydrocarbon homopolymer of length seven ($n = 7$). It provides a simple, tractable problem for which the conformational space can (almost) be exhaustively explored. While this structure reflects great simplifications over the general molecular conformation problem, its complexity must not be underestimated.

The potential energy function is given by (1), where $r_{i,l}$ is the distance between beads, in this case hydrocarbons, separated by two or more residues along the central chain, $k_l = 310.0$ kcal/Ang^2, $k_\theta = 40.0$ kcal/rad^2, $V_n = 1.3$ kcal, $\varepsilon_{i,l} = 0.181$ kcal, $\sigma_{i,l} = 4.0$ Ang, $l_0 = 1.526$ Ang, $\theta_0 = 109.47^\circ$ (1.91 rad), and $H_{i,l} = 1$ for all pairs i,l. Notice that l_0 and θ_0 represent the preferred bond angles and bond lengths and $\bar{n}$ is set to three, providing three preferred torsion angles at 60°, 180°, and 300°, while $\delta=0^\circ$. Since all hydrocarbons can be considered identical, only one value of the $\varepsilon_{i,l}$ and $\sigma_{i,l}$ is required for all i and l.

Despite these simplifications, the conformation problem remains very difficult. Since torsion angles have three preferred positions (60°,180°,300°), 3^{n-3}, or 81, local minima are expected (recall that three of the torsion angles are fixed). Tests on the potential function support this estimate: by starting from 1000 random starting points, 77 local minima were discovered. However, merely selecting preferred angles does not guarantee that the initial conformer is also a local minimum. Notice in Table 5.1 that the local minimum found by

Table 5.1 Selected Minima of Heptane

$U(\phi)$	Initial Position, $\phi^{(0)}$				Minimized Position, ϕ^*			
-0.339727	180	180	180	180	180.000000	180.000000	180.000000	180.000000
-0.215609	-60	60	-60	60	-62.657878	179.600918	-179.857154	179.956953
-0.215524	-60	180	180	180	-62.706316	179.627701	180.119720	180.000000
-0.104118	60	60	-60	60	62.550498	180.191468	-179.798849	62.592457
2.097992	180	-60	60	60	179.378780	-79.278332	79.236470	62.423103
4.384072	180	-60	60	-60	178.741629	-93.723491	69.284953	-94.158124

starting at (-60,60,-60,60) results in the same local minimum as the point with initial torsion angles (-60,180,180,180). The global minimum for heptane corresponds to the point with all torsion angles equal to 180 (see Table 5.1 and Figure 5.1).

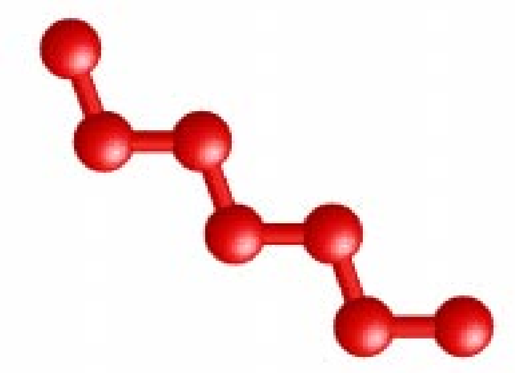

Figure 5.1 Global Minimum Conformation, ϕ^*, for Heptane

While it may seem that 77 local minima is relatively small, larger *n*-chain hydrocarbons produce an exponentially increasing number (i.e. $O(3^n)$) of local minima. A molecule with as few as 12 beads has 19,683 possible preferred angle configurations, with most of these likely to produce a local minimum. It can clearly be seen that finding the global minimum for chains of even moderate length (e.g. $n = 20$ or more) is intractable with exhaustive methods. The next few sections discuss a method which attempts to find the global solution without resorting to an exponential dependence on problem size.

6. A Convex Global Underestimator

One possibility for aiding the search for the global minimum of the molecule's potential energy function is to use a global underestimator to localize the search in the region of the global minimum. This new method is designed to fit all known local minima with a convex function which underestimates all of them, but which differs from them by the minimum possible amount in the discrete L_1 norm. The minimum of this underestimator is used to predict the global minimum for the function, allowing a localized conformer search to be performed based on the predicted minimum. A new set of conformers generated by the localized search can then serve as a basis for another quadratic underestimation. After several repetitions, the global minimum can be found with reasonable assurance.

The use of an underestimating function allows the translation of a *very* complex function into a simple underestimator. If the underestimator is well suited to the problem (i.e. provides accurate predictions for the global minimum), immense savings in time can be achieved. The presence of quadratic terms in the length and bond angle portions of the molecular energy function support the use of a convex quadratic to provide a suitable approximation for this problem.

This method is presented in terms of a differentiable function $F(\phi)$, where $\phi \in \mathbf{R}^{\tau}$ (τ represents the number of torsion angles), and where $F(\phi)$ has many local minima. Since the bond lengths and bond angle terms in the potential function carry severe quadratic penalties, they can be assumed to be fixed for the purposes of the global underestimator. Therefore, ϕ is a vector of torsion angles of length τ, where $\tau = n - 3$ and n is the number of residues (recall that the first three are fixed by definition). To begin the iterative process, a set of $k \geq 2\tau+1$ distinct local minima are computed. This can be done with relative ease by using an efficient unconstrained minimizer, starting with a large enough set of points chosen at random in an initial hyperrectangle $H\phi$, which is assumed to enclose the torsion angle space.

Assuming that $k \geq 2\tau+1$ local minima $\phi^{(j)}$, for j=1,...,k, have been computed, a convex quadratic underestimator function $\Psi(\phi)$ is now fitted to these local minima so that it underestimates all the local minima, and normally interpolates $F(\phi^{(j)})$ at $2\tau+1$ points. This is accomplished by determining the coefficients in the function $\Psi(\phi)$ so that

$$\delta_j = F(\phi^{(j)}) - \Psi(\phi^{(j)}) \geq 0 \tag{2}$$

for j=1,...,k, and where $\sum_{j=1}^{k} \delta_j$ is minimized. That is, the difference between $F(\phi)$ and $\Psi(\phi)$ is minimized in the discrete L_1 norm over the set of k local minima $\phi^{(j)}$, j=1,...,k. The underestimating function $\Psi(\phi)$ is given by

$$\Psi(\phi) = c_0 + \sum_{i=1}^{\tau} \left(c_i \phi_i + \frac{1}{2} d_i \phi_i^2 \right). \tag{3}$$

Note that c_i and d_i appear linearly in the constraints of (2) for each local minimum $\phi^{(j)}$. Convexity of this quadratic function is guaranteed by requiring that $d_i \geq 0$ for i=1,...,τ. Other linear combinations of convex functions could also be used, but this quadratic function is the simplest.

Additionally, in order to guarantee that $\Psi(\phi)$ attains its global minimum Ψ_{min} in the hyperrectangle $H\phi = \{\phi_i : 0 \leq \underline{\phi}_i \leq \phi_i \leq \phi_i \leq 2\pi\}$, the additional set of constraints are imposed on the coefficients of $\Psi(\phi)$:

$$c_i + \underline{\phi}_i d_i \leq 0 \text{ and } c_i + \bar{\phi}_i d_i \geq 0 \text{ for } i=1,\ldots,\tau. \tag{4}$$

Note that the satisfaction of (4) implies that $c_i \leq 0$ and $d_i \geq 0$ for i=1,...,τ.

7. Constructing the Underestimating Function

The unknown coefficients c_i, i=0,...,τ, and d_i, i=1,...,τ, can be determined by a linear program which may be considered to be in the dual form. For reasons of efficiency, the equivalent primal of this problem is actually solved, as described below. The solution to this primal linear program provides an optimal dual vector, which immediately gives the underestimating function coefficients c_i and d_i. Since the convex quadratic function $\Psi(\phi)$ gives a global approximation to the local minima of $F(\phi)$, then its easily computed global minimum function value $\Psi_{\min}$ is a good candidate for an approximation to the global minimum of the correct energy function $F(\phi)$.

An efficient linear programming formulation and solution satisfying (2), (3), and (4) will now be summarized. Let $f^{(j)} = F(\phi^{(j)})$, for j=1,...,k, and let $f \in \mathbf{R}^k$ be the vector with elements $f^{(j)}$. Also let $\omega^{(j)} \in \mathbf{R}^\tau$ be the vector with elements $1/2(\phi^{(j)}{}_i)^2$, i=1,...,τ, and $e_k \in \mathbf{R}^k$ be the vector of ones. Now define the following two matrices $\Phi \in \mathbf{R}^{(\tau+1)\times k}$ and $\Omega \in \mathbf{R}^{\tau\times k}$:

$$\text{(5)} \qquad \Phi = \begin{bmatrix} e_k^T \\ \phi^{(1)}\,\phi^{(2)}\ldots\phi^{(k)} \end{bmatrix},\ \Omega = \begin{bmatrix} \omega^{(1)}\,\omega^{(2)}\ldots\omega^{(k)} \end{bmatrix}.$$

Finally, let $c \in \mathbf{R}^{\tau+1}$, $d \in \mathbf{R}^\tau$, and $\delta \in \mathbf{R}^k$ be the vectors with elements c_i, d_i, and δ_i, respectively. Then (2), (3), and (4) can be restated as the linear program (with variables c, d, and δ):

$$\underset{\text{c, d, }\delta}{\text{minimize}} \quad e_k^T\delta$$

$$\text{(6)} \qquad \text{subject to} \begin{bmatrix} \Phi^T & \Omega^T & 0 \\ -\Phi^T & -\Omega^T & -I_k \\ I'_\tau & \underline{D} & 0 \\ -I'_\tau & -\overline{D} & 0 \end{bmatrix} \begin{bmatrix} c \\ d \\ \delta \end{bmatrix} \le \begin{bmatrix} f \\ -f \\ 0 \\ 0 \end{bmatrix}$$

where $\underline{D} = diag(\underline{\phi}_1,\underline{\phi}_2,\ldots,\underline{\phi}_\tau)$, $\overline{D} = diag(\overline{\phi}_1,\overline{\phi}_2,\ldots,\overline{\phi}_\tau)$, I_k is the identity matrix of order k, and I'_τ is the τx(τ+1) "augmented" matrix $[0 : I_\tau]$ where I_τ is the identity matrix of order τ.

Since the matrix in (6) has more rows than columns $2\cdot(k+\tau)$ rows and $k+2\tau+1$ columns, where $k \ge 2\tau+1$), it is computationally more efficient to consider it as a dual problem, and to solve the equivalent primal. After some simple transformations, this primal problem reduces to:

$$\underset{y_1, y_2, y_3}{\text{minimize}} \quad f^Ty_1 - f^Te_k$$

$$\text{(7)} \qquad \text{subject to} \begin{bmatrix} \Phi & I'^T_\tau & -I'^T_\tau \\ \Omega & \underline{D} & -\overline{D} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \Phi e_k \\ \Omega e_k \end{bmatrix}, \quad y_1, y_2, y_3 \ge 0$$

which has only $2\tau+1$ rows and $k+2\tau \ge 4\tau+1$ columns, and the obvious initial feasible solution $y_1 = e_k$ and $y_2 = y_3 = 0$. Furthermore, since the first of the $2\tau+1$ constraints in (7) in fact requires that $e_k{}^Ty_1 = 1$, then the function $f^Ty_1 - f^Te_k$ is also bounded below, and so this primal linear program always has an optimal solution. This optimal solution gives the values of c, d, and δ via the dual vectors, and also determines which values of $f^{(j)}$ are interpolated by the potential function $\Psi(\phi)$. That is, the basic columns in the optimal solution to (7) correspond to the conformations $\phi^{(j)}$ for which $f^{(j)} = \Psi(\phi^{(j)})$.

Note that once an optimal solution to (7) has been obtained, the addition of new local minima is very easy. It is done by simply adding new columns to Φ and Ω, and therefore to the constraint matrix in (7). The number of primal rows remains fixed at $2\tau+1$, independent of the number k of local minima.

8. Iterative Approximation to the Global Minimum

The convex quadratic underestimating function $\Psi(\phi)$ determined by the values $c \in \mathbf{R}^{\tau+1}$ and $d \in \mathbf{R}^{\tau}$ now provides a global approximation to the local minima of $F(\phi)$, and its easily computed global minimum point $\phi_{\min}$ is given by $(\phi_{\min})_i = -c_i/d_i$, i=1,...,τ, with corresponding function value $\Psi_{\min}$ given by $\Psi_{\min} = c_0 - \sum_{i=1}^{\tau} c_i^2 / (2d_i)$. The value $\Psi_{\min}$ is a good candidate for an approximation to the global minimum of the correct energy function $F(\phi)$, and so $\phi_{\min}$ can be used as an initial starting point around which additional configurations (i.e. local minima) should be generated. These local minima are added to the constraint matrix in (7) and the process is repeated. Before each iteration of this process, it is necessary to reduce the volume of the hyperrectangle $H\phi$ over which the new configurations are produced so that a tighter fit of $\Psi(\phi)$ to the local minima "near" $\phi_{\min}$ is constructed.

The rate and method by which the hyperrectangle size is decreased, and the number of additional local minima computed at each iteration will be determined by computational testing. But clearly the method depends most heavily on computing local minima quickly and on solving the resulting linear program efficiently to determine the approximating function $\Psi(\phi)$ over the current hyperrectangle.

If E_c is a cutoff energy, then one means for decreasing the size of the hyperrectangle $H\phi$ at any step is to let $H\phi = \{\phi\colon \Psi(\phi) \le E_c\}$. To get the bounds of $H\phi$, consider $\Psi(\phi) \le E_c$ where $\Psi(\phi)$ satisfies (3). Then limiting ϕ_i requires that

$$\sum_{i=1}^{\tau} \left(c_i\phi_i + \frac{1}{2}d_i\phi_i^2 \right) \le E_c - c_0 .$$

As before, the minimum value of $\Psi(\phi)$ is attained when $\phi_i = -c_i/d_i$, i=1,...,τ. Assigning this minimum value to each ϕ_i, except ϕ_k, then results in

$$c_k\phi_k + \frac{1}{2}d_k\phi_k^2 \le E_c - c_0 + \frac{1}{2}\sum_{i \ne k} c_i^2 / d_i \equiv \beta_k .$$

The lower and upper bounds on ϕ_k, k=1,...,τ, are given by the roots of the quadratic equation

$$c_k\phi_k + \frac{1}{2}d_k\phi_k^2 \ = \ \beta_k . \tag{8}$$

Hence, these bounds can be used to define the new hyperrectangle $H\phi$ in which to generate new configurations.

Clearly, if E_c is reduced, the size of $H\phi$ is also reduced. At every iteration the predicted global minimum value $\Psi_{\min}$ satisfies $\Psi_{\min} \le F(\phi^*)$, where ϕ^* is the smallest known local minimum conformation. Therefore, $E_c = F(\phi^*)$ is often a good choice. If at least one improved point ϕ, with $F(\phi) < F(\phi^*)$, is obtained in each iteration, then the search domain $H\phi$ will strictly decrease at each iteration, and may decrease substantially in some iterations.

9. The Algorithm

Based on the results of the previous three sections, a general method for computing the global, or near global, energy minimum of the potential energy function U can now be described. Recall that $\phi \in \mathbf{R}^{\tau}$ where τ represents the number of torsion angles, and the bond lengths and bond angle are assumed to be fixed for the purposes of the global underestimator.

1. Compute $k \geq 2\tau+1$ distinct local minima $\phi^{(j)}$, for $j=1,...,k$, of the function $F(\phi)$.
2. Compute the convex quadratic underestimator function

$$\Psi(\phi) = c_0 + \sum_{i=1}^{\tau} \left(c_i \phi_i + \frac{1}{2} d_i \phi_i^2 \right)$$

by solving the linear program

$$\underset{y_1, y_2, y_3}{\text{minimize}} \quad f^T y_1 - f^T e_k$$

$$\text{subject to} \quad \begin{bmatrix} \Phi & I'^T_{\tau} & -I'^T_{\tau} \\ \Omega & \underline{D} & -\overline{D} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \Phi e_k \\ \Omega e_k \end{bmatrix}, \quad y_1, y_2, y_3 \geq 0$$

The optimal solution to this linear program gives the values of c and d via the dual vectors.

3. Compute the predicted global minimum point ϕ_{min} given by $(\phi_{\text{min}})_i = -c_i/d_i$, $i=1,...,\tau$, with corresponding function value Ψ_{min} given by $\Psi_{\text{min}} = c_0 - \sum_{i=1}^{\tau} c_i^2 / (2d_i)$.
4. If $\phi_{\text{min}} = \phi^*$, where $\phi^* = \text{argmin}\{F(\phi^{(j)}), j=1,2,... \}$ is the best local minimum found so far, then stop and report ϕ^* as the approximate global minimum conformation.
5. Reduce the volume of the hyperrectangle Hϕ over which the new configurations will be produced, and remove all columns from Φ and Ω which correspond to the conformations which are excluded from Hϕ.
6. Use ϕ_{min} as an initial starting point around which additional local minima $\phi^{(j)}$ of $F(\phi)$ (restricted to the hyperrectangle Hϕ) are generated. Add these new local minimum conformations as columns to the matrices Φ and Ω.
7. Return to step 2.

The number of new local minima to be generated in step 6 is unspecified since there is currently no theory to guide this choice. In general, a value exceeding $2\tau+1$ would be required for the construction of another convex quadratic underestimator in the next iteration (step 2). In addition, the means by which the volume of the hyperrectangle Hϕ is reduced in step 5 may vary. One such method, presented in section 8, would use the two roots of (8) to define the new bounds of Hϕ. Another method would be simply to use Hϕ = $\{\phi_i : (\phi_{\text{min}})_i - \delta_i \leq \phi_i \leq (\phi_{\text{min}})_i + \delta_i, i=1,...,\tau\}$ where $\delta_i = |(\phi_{\text{min}})_i - (\phi^*)_i|$, $i=1,...,\tau$.

10. Computational Results

The computational results presented in this section were obtained on the Cray Y-MP C90 and Cray T3D systems (at the Minnesota Supercomputer Institute) using the network PVM message passing system for communication between the C90 and T3D. Steps 1 and 6 of the algorithm (presented in section 9) were performed in parallel on 8 processors of the T3D, while the remaining steps of the algorithm were done sequentially on the C90.

Using this system, solutions were obtained for the *n*-chain hydrocarbon problem with lengths 4 through 22, 25, and 30. As an example, a solution for the 7 bead hydrocarbon heptane was obtained in only 8 seconds after 3 iterations (a total of 180 local minimizations were performed). Table 10.1 summarizes the minimization of the potential function for heptane (an entry of NC indicates “no change” between the predicted and computed torsion angles in that step). Notice that the first iteration reports the current “best” conformation (corresponding to the lowest potential energy value) to be approximately the set of torsion angles (180˚,180˚,300˚,180˚). Recall that the bond lengths and bond angles are considered fixed for the purposes of the global underestimator. However, the global underestimator predicts a conformation close to (180˚, 180˚, 180˚, 180˚). After the second iteration, with Hϕ defined by {175.45˚ ≤ ϕ_1 ≤ 215.45˚, 172.96˚ ≤ ϕ_2 ≤ 203.96˚, 23.06˚ ≤ ϕ_3 ≤ 303.06˚, 175.98˚ ≤ ϕ_4 ≤ 215.98˚}, we find that a better function value can be realized at this conformation. The underestimator and the current global minimum now reside at almost identical points. The third iteration confirms (180˚, 180˚, 180˚, 180˚) as the global minimum. This result has been shown through enumeration to be the correct global minimum conformation for the given potential energy function (the solution is shown in Figure 5.1).

Table 10.1 Solution Summary for Heptane

	Initial Position, $\phi^{(0)}$	Randomize over Radius	New Global Minimum (25 random points)	Predicted Global Minimum	Randomize over Radius	New Global Minimum (20 random points)	Predicted Global Minimum	Randomize over Radius	New Global Minimum (10 random points)	Predicted Global Minimum
φ_1	20	180	180.12	195.45	20	180.01	NC	4	180.02	NC
φ_2	330	180	179.53	187.96	15	179.99	NC	4	179.99	179.88
φ_3	278	180	297.50	163.06	140	180.06	180.14	5	180.12	NC
φ_4	-60	180	179.63	195.98	20	179.97	179.88	5	179.91	NC
F(ϕ)			-.2967			-.3398			-.3398	
									ϕ*	

As the problem size *n* increases, the number of local minima increases exponentially at the rate of $O(3^n)$ as can be seen in Figure 10.1. With exponential increases in the number of local minima, exhaustive searches for the global minimum quickly become computationally intractable. The time required for exhaustive techniques is driven by the number of local minimizations needed to generate and “relax” each possible conformation. For example, there are 3^{12} = 531,441 possible conformations for the 15 bead problem, and hence, to test every possible local minimum conformation is expected to take over 100 hours on a Cray Y-MP C90!

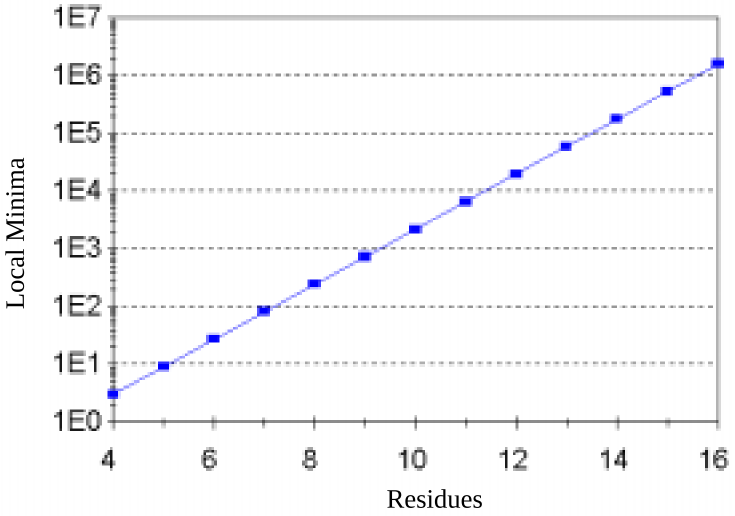

Figure 10.1 Total Number of Local Minima as a Function of the Number of Residues

Through the use of the global underestimator we can avoid such prohibitive numbers of calculations; however, the underestimator still remains dependent on the accumulation of a large number of local minima. In contrast to enumeration, the underestimator method requires a much less dramatic increase in the number of computations. Since the total number of underestimation steps needed to find a global minimum remains close to constant (observed experimentally to be less than 10), and each step requires $O(n)$ local minima to form the underestimator, a linear increase in number of local minimizations is expected. In addition, since every function evaluation requires $O(n^2)$ time while every gradient evaluation requires $O(n^3)$ time, due to the Lennard-Jones term, the underestimator method produces a final prediction in $O(n^4)$ time on average. The effect of increases in problem size on the total time for computing the global minimum can be seen in Figure 10.2. The time required to compute the global minimum as a function of the number of beads n can be approximated (using a least squares fit) by the function $\mathrm{T}(n) = 0.1311\ n^4 - 5.843\ n^3 + 94.82\ n^2 - 623.9\ n + 1377$ seconds, and therefore increases at an average rate of $\mathrm{O}(n^4)$ as expected. In fact, the 22 bead problem took 36 minutes for the final prediction (Figure 10.3). This prediction was made with nine iterations of the global underestimator, and a total of 1950 local minimizations were performed. Further increases in problem size will face even larger increases in execution time: a 50 bead problem is roughly predicted to require 8 hours on the Cray Y-MP C90 / Cray T3D system (using all 64 processors). The global energy minimum conformations with corresponding global minimum function values, for $n = 16$ through $n = 22$, are given in Appendix B (Section 13).

As an alternate means for observing the time complexity of this algorithm, Figure 10.4 shows the solution time *per iteration* required to obtain the global energy minimum. That is, the total solution time is normalized by the number of iterations. Since the number of iterations has been observed to be less than 10 in all cases, the time complexity in this case is also $\mathrm{O}(n^4)$. In this case, the function $\overline{\mathrm{T}}(n) = 0.00880\ n^4 - 0.361\ n^3 + 5.66\ n^2 - 36.4\ n + 80.7$ seconds approximates this curve.

Finally, it must be noted that although this method is not *guaranteed* to find the global minimum conformation of the potential function U, it tends to perform reasonably well on

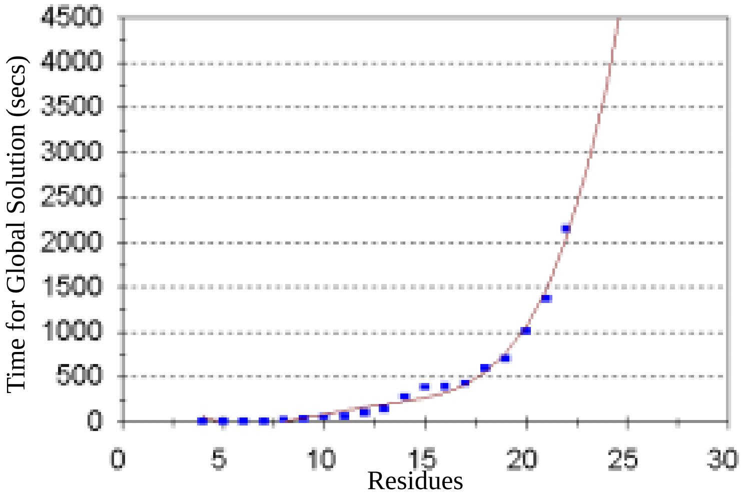

Figure 10.2 Total CPU Time for Global Minimum Computation as a Function of the Number of Residues

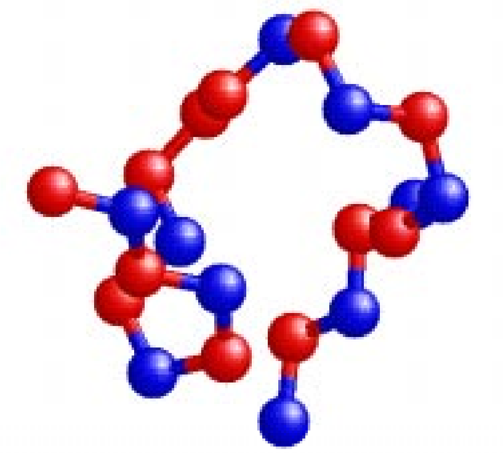

Figure 10.3 Global Energy Minimum for 22 Bead Hydrocarbon

problems of moderate size ($n \leq 30$). In fact, based on the computational results obtained for $n = 4$ though $n = 22$, the computed global minimum energy for the potential function U, as a function of the number of beads n, can be approximated (using a least squares fit) by the function $E(n) = 1.24 - 0.118\, n - 0.0202\, n^2$. Figure 10.5 shows this function superimposed on the computed global minima obtained by the algorithm of section 9 for all hydrocarbon chains of size $n \leq 22$ and also for $n = 25$ and $n = 30$. Notice that the function $E(n)$ does indeed correctly predict the global minimum energy for $n = 25$ and $n = 30$, the two largest problems. This important property permits estimation of the global minimum energy for larger molecular chains, and can also be used to accelerate the global minimization algorithm. Accordingly, the predicted global minimum energy for an $n = 50$ bead conformation is expected to obtain a global minimum energy of approximately -55.14 kcal. If

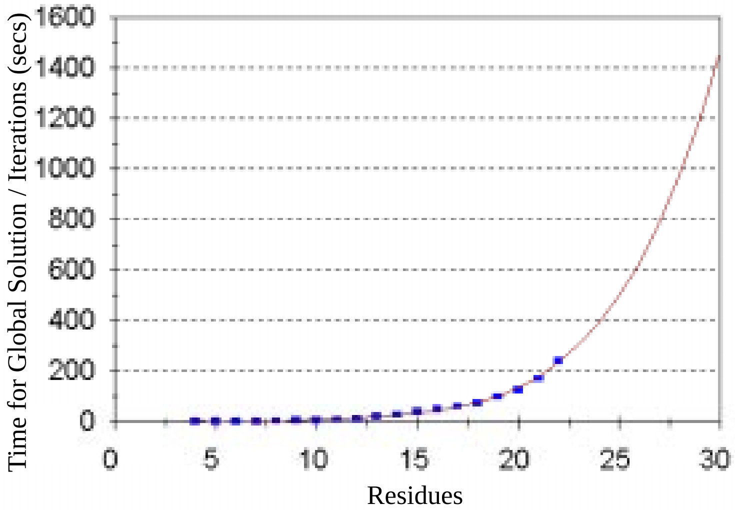

Figure 10.4 Total CPU Time for Global Minimum Computation per Iteration as a Function of the Number of Residues

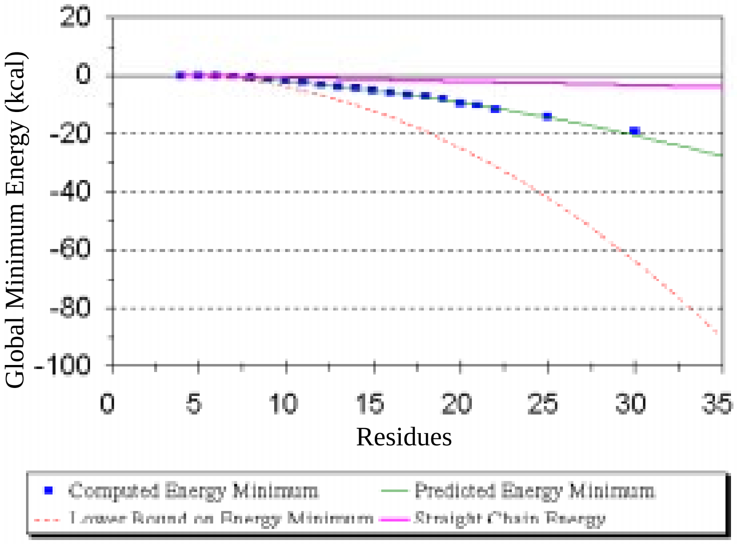

Figure 10.5 Computed Global Energy Minima as a Function of the Number of Residues

we define $f_0 = -\sum \varepsilon_{i,\,l}$, where the sum is computed over all pairs (a_i,a_l) of beads separated by more than two along the primary chain, then f_0 represents a lower bound on the potential function U from (1). Figure 10.5 also shows f_0 (for n-chain hydrocarbons), for comparison with the computed energy minima, which is given by $f_0(n)$ = -1.05855 + 0.62445 n - 0.0905 n^2, and the straight chain potential energy value (all torsions ϕ_i = 180˚) which can be regarded as a simple upper bound. Finally, it should be emphasized that while the minimum energy is a smooth function of the chain length, the corresponding minimum energy conformations (as illustrated in Appendix B) may differ substantailly as the chain length increases.

11. Conclusions

A convex quadratic global underestimator method was used to predict stable molecular conformations for *n*-chain homopolymers (hydrocarbons) of length 4 through 22, 25, and 30. For the convex global underestimator approach to realize its full potential, not only must massively parallel machines be used effectively to reduce the time required to compute the large number of local minima, but also, since the local minimization procedures rely on many function evaluations, improved methods for computing the potential function must be designed. Such improvements in the function evaluation time (including the conversions) would be of tremendous benefit, and perhaps will provide the key to reducing the local minimization times. Recent work in [6] has attempted to specifically address this problem.

12. Appendix A: Transformations and Analytic Derivatives

If the three dimensional cartesian position of bead a_m, m=1,...,n, is represented by (x_m, y_m, z_m) $\in R^3$, then the following sequence of matrix products is sufficient to convert from internal molecular coordinates to the cartesian representation, and is based solely on the internal molecular coordinates of the previous m-1 beads in the chain:

$$\begin{bmatrix} x_m \\ y_m \\ z_m \\ 1 \end{bmatrix} = B_1 B_2 \ldots B_m \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{9}$$

where the transformation matrices $B_i \in R^{4x4}$, i=1,...,m, were first defined in [11] to be of the form

$$B_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B_2 = \begin{bmatrix} -1 & 0 & 0 & -l_2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$B_i = \begin{bmatrix} -\cos\theta_i & -\sin\theta_i & 0 & -l_i\cos\theta_i \\ \sin\theta_i\cos\varphi_i & -\cos\theta_i\cos\varphi_i & -\sin\varphi_i & l_i\sin\theta_i\cos\varphi_i \\ \sin\theta_i\cos\varphi_i & -\cos\theta_i\sin\varphi_i & \cos\varphi_i & l_i\sin\theta_i\sin\varphi_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

for i=3,...,m. Hence, in order to compute the cartesian positions of each of the n beads a_1, a_2, ..., a_n, an O(n) algorithm involving only 4x4 matrix products is sufficient.

For an n bead molecular structure, the potential function U depends on the 3n - 6 independent variables l_k, k=2,...,n, θ_k, k=3,...,n, and φ_k, k=4,...,n. Computing the gradient of the function U, with respect to the internal coordinates, will therefore require an analytic formulation for $\partial U/\partial\varphi_k$, $\partial U/\partial\theta_k$, and $\partial U/\partial l_k$. Each of these can be computed using a sequence of matrix products in a manner similar to the transformations above. To compute each, first notice from (1) that

$$\frac{\partial U}{\partial \varphi_k} = -\frac{\bar{n} V_{\bar{n}}}{2} \sin(\bar{n}\varphi_k + \delta) + \sum_{i,l} \frac{12\varepsilon_{i,l}}{r_{i,l}} \left(\left(\frac{\sigma_{i,l}}{r_{i,l}} \right)^{12} + \left(\frac{\sigma_{i,l}}{r_{i,l}} \right)^{6} \right) \frac{\partial r_{i,l}}{\partial \varphi_k} \tag{11}$$

$$\frac{\partial U}{\partial \theta_k} = k_\theta (\theta_k - \theta_0) + \sum_{i,l} \frac{12\varepsilon_{i,l}}{r_{i,l}} \left(\left(\frac{\sigma_{i,l}}{r_{i,l}} \right)^{12} + \left(\frac{\sigma_{i,l}}{r_{i,l}} \right)^{6} \right) \frac{\partial r_{i,l}}{\partial \theta_k} \tag{12}$$

$$\frac{\partial U}{\partial l_k} = k_l (l_k - l_0) + \sum_{i,l} \frac{12\varepsilon_{i,l}}{r_{i,l}} \left(\left(\frac{\sigma_{i,l}}{r_{i,l}} \right)^{12} + \left(\frac{\sigma_{i,l}}{r_{i,l}} \right)^{6} \right) \frac{\partial r_{i,l}}{\partial l_k} . \tag{13}$$

The dependence of $r_{i,l}$ on φ_k, θ_k, and l_k is given by the sequence of 4x4 matrix products in (9). To analytically compute the terms $\partial r_{i,l}/\partial \varphi_k$, $\partial r_{i,l}/\partial \theta_k$, and $\partial r_{i,l}/\partial l_k$, first define $w_m = (x_m, y_m, z_m, 1)^T$ and notice that $r_{i,l}^2 = (w_i - w_l)^T(w_i - w_l) = v_{i,l}^T v_{i,l}$ where

$$v_{i,l} = w_i - w_l = \left(\prod_{j=1}^{i} B_j \right) \left(I - \prod_{j=i+1}^{l} B_j \right) e_4 \tag{14}$$

with $e_4 = (0,0,0,1)^T$. Letting

$$B_{i,l} = \left(\prod_{j=1}^{i} B_j \right) \left(I - \prod_{j=i+1}^{l} B_j \right) \tag{15}$$

then $v_{i,l} = B_{i,l} e_4$ and $r_{i,l}^2 = (B_{i,l} e_4)^T (B_{i,l} e_4)$. Therefore,

$$\frac{\partial r_{i,l}}{\partial \varphi_k} = \frac{1}{2 r_{i,l}} \frac{\partial r_{i,l}^2}{\partial \varphi_k} = \frac{e_4^T B_{i,l}^T}{r_{i,l}} \left(\frac{\partial B_{i,l}}{\partial \varphi_k} \right) e_4$$

and similarly,

$$\frac{\partial r_{i,l}}{\partial \theta_k} = \frac{e_4^T B_{i,l}^T}{r_{i,l}} \left(\frac{\partial B_{i,l}}{\partial \theta_k} \right) e_4$$

$$\frac{\partial r_{i,l}}{\partial l_k} = \frac{e_4^T B_{i,l}^T}{r_{i,l}} \left(\frac{\partial B_{i,l}}{\partial l_k} \right) e_4 .$$

Finally, the dependence of $B_{i,l}$ on φ_k ($k = 4,...,n$), θ_k ($k = 3,...,n$), and l_k ($k = 2,...,n$) can be expressed by the following three cases

$$\frac{\partial B_{i,l}}{\partial \alpha_k} = \begin{cases} \underline{0} & l < k \\ -\left(\prod_{j=1}^{k-1} B_j \right) \frac{\partial B_k}{\partial \alpha_k} \left(\prod_{j=k+1}^{l} B_j \right) & i < k \le l \\ \left(\prod_{j=1}^{k-1} B_j \right) \frac{\partial B_k}{\partial \alpha_k} \left(\prod_{j=k+1}^{i} B_j \right) \left(I - \prod_{j=i+1}^{l} B_j \right) & 1 \le k \le i \end{cases} \tag{16}$$

where α_k represents any one of the three parameters φ_k, θ_k, or l_k, and $\underline{0}$ is the 4x4 zero matrix. Note that whereas computing the potential function U requires $O(n^2)$ steps (due to the pairwise Lennard-Jones terms), the analytic evaluation of the gradient requires $O(n^4)$ steps.

13. Appendix B: Minimum Energy Conformations for *n* = 16 - 22.

Table 13.1

Global Minimum Conformation						
n = 16	*n* = 17	*n* = 18	*n* = 19	*n* = 20	*n* = 21	*n* = 22
173°	172°	181°	296°	175°	295°	180°
188°	189°	176°	180°	65°	181°	182°
63°	62°	293°	172°	195°	177°	295°
59°	59°	292°	296°	67°	291°	159°
187°	186°	165°	292°	65°	180°	291°
60°	60°	294°	152°	195°	290°	296°
63°	64°	193°	290°	63°	292°	172°
60°	60°	166°	172°	174°	171°	295°
187°	186°	61°	178°	177°	292°	183°
59°	58°	197°	63°	56°	298°	181°
63°	62°	65°	65°	184°	297°	298°
188°	192°	66°	178°	61°	160°	177°
173°	168°	193°	67°	64°	299°	297°
	185°	67°	67°	186°	181°	288°
		181°	173°	62°	183°	152°
			182°	69°	65°	293°
				185°	181°	177°
					182°	175°
						59°
Global Minimum Potential Energy						
-5.783	**-6.397**	**-7.634**	**-8.025**	**-9.766**	**-10.44**	**-11.81**

14. References

1. R.A. Abagyan, *Towards Protein Folding by Global Energy Optimization*, Federation of European Biochemical Societies: Letters **325** (1993), 17-22.
2. S.A. Benner and D.L. Gerloff, *Predicting the Conformation of Proteins: Man versus Machine,* Federation of European Biochemical Societies: Letters **325** (1993), 29-33.
3. D.A. Case, *Computer Simulations of Protein Dynamics and Thermodynamics*, IEEE: Computer Science & Engineering, October (1993), 47-57.
4. K.A. Dill, *Dominant Forces in Protein Folding*, Biochemistry **29** (1990), 7133-7155.
5. D.M. Ferguson, A. Marsh, T. Metzger, D. Garret, and K. Kastella, *Conformational Searches for the Global Minimum of Protein Models,* Journal of Global Optimization **4** (1994), 209-227.
6. L. Greengard, *The Rapid Evaluation of Potential Fields in Particle Systems*, The MIT Press, Cambridge, 1988.

7. A.L. Lehninger, *Biochemistry: The Molecular Basis of Cell Structure and Function*, Worth Publishers, New York, 1970.
8. K. Merz and S. Le Grand, *The Protein Folding Problem and Tertiary Structure Prediction*, Birkhauser, Boston, 1994.
9. F.M. Richards, *The Protein Folding Problem*, Scientific American (1991), 54-63.
10. S. Sun, *Reduced representation model of protein structure prediction: statistical potential and genetic algorithms*, Protein Science **2** (1993), 762-785.
11. H.B. Thompson, *Calculation of Cartesian Coordinates and Their Derivatives from Internal Molecular Coordinates*, The Journal of Chemical Physics **47** (1967), 3407-3410.
12. V.H. Walke, *Computational Solutions to the Protein Folding Problem*, Trident Scholar Report No. 223, U.S. Naval Academy, Annapolis, Maryland (1994).

A.T. PHILLIPS, COMPUTER SCIENCE DEPARTMENT, UNITED STATES NAVAL ACADEMY, ANNAPOLIS, MD 21402

J.B. ROSEN, COMPUTER SCIENCE AND ENGINEERING DEPARTMENT, UNIVERSITY OF CALIFORNIA AT SAN DIEGO, SAN DIEGO, CA 92093

V.H. WALKE, COMPUTER SCIENCE DEPARTMENT, UNIVERISTY OF TEXAS AT AUSTIN, AUSTIN, TX 78712